GRPO

$$\hat{A}_{i,t} = \frac{R(q,o_i) - \text{mean}\{R(q,o_1),\ldots,R(q,o_G)\}}{\text{std}\{R(q,o_1),\ldots,R(q,o_G)\}}$$

$$\mathcal{L}_{\text{GRPO}_{i,t}} = \min\left[\frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}\hat{A}_{i,t},\, clip_{1\pm\epsilon}\left[\frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}\right]\hat{A}_{i,t}\right]$$

GSPO

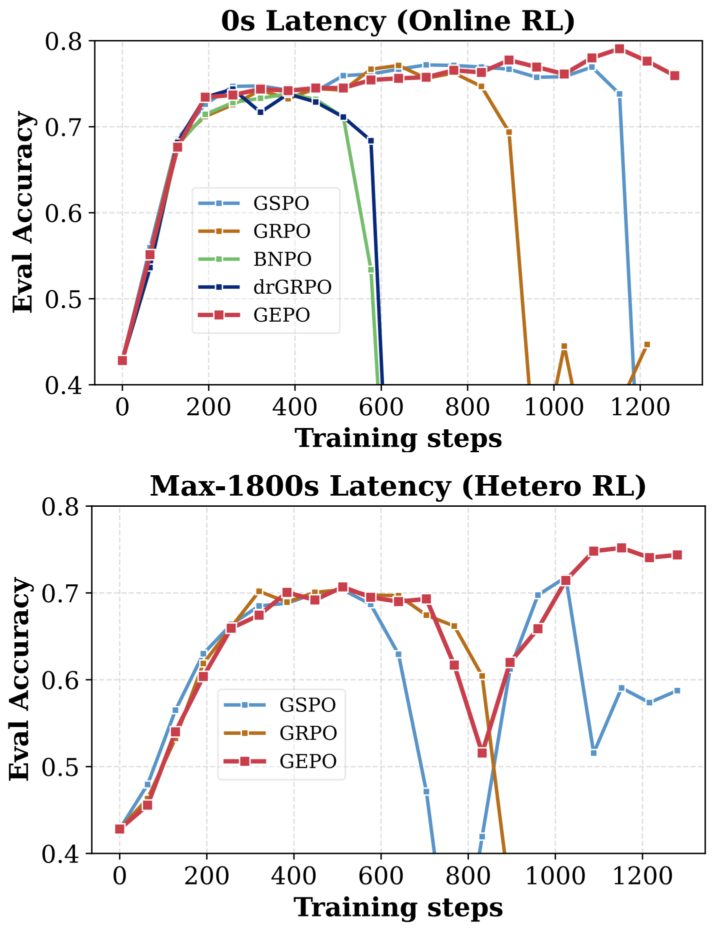
$$\hat{A}_{i} = \frac{R(q,o_i) - \text{mean}\{R(q,o_1),\ldots,R(q,o_G)\}}{\text{std}\{R(q,o_1),\ldots,R(q,o_G)\}}$$

$$\mathcal{L}_{\text{GSPO}_{i}} = \min\left[\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)}\hat{A}_{i},\, clip_{1\pm\varepsilon}\left[\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)}\right]\hat{A}_{i}\right]$$

GEPO

$$\hat{A}_{i} = \frac{R(q,o_i) - \text{mean}\{R(q,o_1),\ldots,R(q,o_G)\}}{\cancel{\text{std}\{R(q,o_1),\ldots,R(q,o_G)\}}}$$

$$\mathcal{L}_{\text{GEPO}_{i}} = \min\left[\underbrace{\frac{\pi_\theta(o_i \mid q)}{\mathbb{E}_{\pi_{\theta_{\text{old}}}(\cdot|q)}\pi_{\theta_{\text{old}}}(o \mid q)}}_{\text{Group Expectation}}\hat{A}_{i},\, clip_{1\pm\varepsilon}\left[\frac{\pi_\theta(o_i \mid q)}{\mathbb{E}_{\pi_{\theta_{\text{old}}}(\cdot|q)}\pi_{\theta_{\text{old}}}(o \mid q)}\right]\hat{A}_{i}\right]$$


0s Latency (Online RL)
Eval Accuracy
Training steps
GSPO
GRPO
BNPO
drGRPO
GEPO
Max-1800s Latency (Hetero RL)
Eval Accuracy
Training steps
GSPO
GRPO
GEPO